Testing Causal Models of Word Meaning in LLMs

Sam Musker (samuel_musker@brown.edu)
Brown University Department of Computer Science, 115 Waterman Street
Providence, RI 02906 USA

Ellie Pavlick (ellie_pavlick@brown.edu)
Brown University Department of Computer Science, 115 Waterman Street
Providence, RI 02906 USA

Abstract

Large Language Models (LLMs) have driven extraordinary improvements in NLP. However, it is unclear how such models represent lexical concepts–i.e., the meanings of the words they use. We evaluate the lexical representations of GPT-4, GPT-3, and Falcon-40B through the lens of HIPE theory, a concept representation theory focused on words describing artifacts (such as "mop", "pencil", and "whistle"). The theory posits a causal graph relating the meanings of such words to the form, use, and history of the referred objects. We test LLMs with the stimuli used by Chaigneau et al. (2004) on human subjects, and consider a variety of prompt designs. Our experiments concern judgements about causal outcomes, object function, and object naming. We do not find clear evidence that GPT-3 or Falcon-40B encode HIPE's causal structure, but find evidence that GPT-4 does. The results contribute to a growing body of research characterizing the representational capacity of LLMs.

Keywords: Large Language Models; Lexical concepts; Causal models

Introduction

The success of large language models (LLMs) at generating human-like text has spurred a wave of recent work which aims to measure the extent to which such models have good representations of word meanings (i.e., lexical concepts). Such work has taken a variety of forms across multiple domains, but in general amounts to measuring the extent to which the conceptual associations encoded by LLMs match human associations. For example, prior work has shown that LLMs correctly associate physical objects with their properties and affordances (Forbes et al., 2019; Da & Kasai, 2019); common nouns with the ontological categories (Da & Kasai, 2019; Ettinger, 2020), and entities with their salient characteristics (Petroni et al., 2019). By and large, the results reported via such studies have been positive, albeit with significant caveats (see Ettinger (2020) and Kassner & Schütze (2020) for specific criticisms and Pavlick (2022) for a general discussion).

Studies like those above are often not framed overtly in theoretical terms. However, implicitly, they assume a theory of lexical concepts in which meaning is defined via a complex network of associations and inferences (Greenberg & Harman, 2005). Such theories are a good first step, but contemporary work in psychology has tended to favor a more nuanced picture, in which lexical concepts are embedded in *causal models* (CMs) of the world (Keil, 1989; Carey, 2009; Sloman, 2005). These CMs can capture complex inferences about word meaning that have been documented in humans–for example, that a raccoon remains one even after it has been surgically altered to look and act like a skunk (Keil, 1989). Such inferences are not easily explained by theories of concepts that rely on naive association or traditional logical entailment.

In this work, we adopt one such causal model theory of lexical concepts, namely the HIPE theory (Chaigneau et al., 2004), and use it to evaluate whether Falcon-40B (Technology Innovation Institute, 2023), GPT-3 (Brown et al., 2020), and GPT-4 (OpenAI, 2023) understand terms referring to basic household objects (specifically, "mop", "pencil", and "whistle"). We test these models on the stimuli which were used to evaluate humans in the original paper. We find that GPT-3 does not track humans in matching the predictions of HIPE theory about the relative importance of factors determining the concepts tested, even when the experiment is repeated in multiple different variations to guard against a false negative. We similarly fail to observe Falcon-40B replicating HIPE theory's predictions. Contrastingly, we find that GPT-4 tracks humans very well in matching the predictions of HIPE theory on a natural reimplementation of the experiment without introducing experiment variations that would increase the chance of the model's success.

This cognitive science-inspired experiment may contribute towards interpreting the representations employed by LLMs. Moreover, our findings raise important questions about how to evaluate conceptual representations in LLMs. In particular, situating our results within a large literature of treating language models as "psycholinguistic subjects" (Futrell et al., 2019), a pertinent question is how to interpret the (increasingly positive) results of LLMs on tests designed to assess humans. If we are hesitant to read success on such tests alone as evidence of "human-like" processing (as we the authors are in this case)–what additional testing do we require?

Related Work

This work contributes to a large body of work on analyzing LLMs as "psycholinguistic subjects" (Futrell et al., 2019) by evaluating their performance on tasks designed to probe human language understanding (Marvin & Linzen, 2018; Warstadt et al., 2020; Ettinger, 2020), and more generally to work that uses counterfactual manipulations of model inputs

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

in order to understand model representations (P.-S. Huang et al., 2020; Goyal et al., 2019). The phenomena we study relate to past work on "commonsense" physical knowledge in LLMs (Bisk et al., 2020; Forbes et al., 2019; W. Huang et al., 2022), but differs in that we are analyzing an LLM through the lens of a particular, empirically-validated theory about conceptual representations in humans.

The HIPE Theory

The HIPE theory (Chaigneau et al., 2004) aims to explain humans' representations of artifacts (in particular, the work uses the common household objects mops, pencils, and whistles). The HIPE theory posits that humans model an artifact using a causal model (CM) involving the artifact's design history (H), the intentions of relevant agents (I), the object's physical structure (P), and events that occur during its use such as actions taken with it (E). More specifically, the theory posits a particular CM as underlying human reasoning about artifacts (Fig. 1). It hypothesizes that the object's design history and the user's goal are distal causes in the CM, while the object's physical structure and the user's actions with respect to it are proximal causes in the CM. Thus, HIPE predicts that, for example, both the physical structure of an object (e.g., having a handle and something absorbent on one end) as well as the reason the object was originally created (e.g., for wiping up water) should affect how appropriate it is to call the object a "mop", but that the latter should have a minimal effect when the former is fully specified.

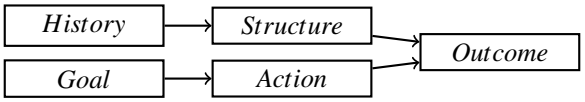

Figure 1: The CM hypothesized by HIPE theory as underlying human representations of artifacts.

Chaigneau et al. (2004) experimentally confirm that such a CM specifies the structure of human representations of artifacts. They construct scenarios describing the history, structure, goal, and action of the three objects. Each scenario is either a baseline scenario in which all four factors are as one would expect them to be, or a compromised scenario in which exactly one of the factors is altered to a compromised description (Figure 2). The subjects are then asked to respond to questions about the object's naming ("Is it appropriate to call this object a mop?"), function ("Does this scenario illustrate the function of a mop?"), or causal outcomes ("Is it likely that, as a result of the events described above, John wiped up the water spill?") using a 1-7 Likert scale. The authors verify that compromising the action has a more pronounced effect than compromising the goal, and likewise that compromising the structure has a more pronounced effect than compromising the design history. This supports the CM's designation of action and structure as proximal causes due to "screening off" (Park & Sloman, 2016).

One day Jane wanted to wipe up a water spill on the kitchen floor, but she didn't have anything to do it with. So she decided to make something. She looked around the house for things that would allow her to make an object for wiping up a water spill on the kitchen floor. She gathered all the materials and made it. When she finished, she left it in the kitchen so she could use it later. *The object consisted of a bundle of thick cloth attached to a 4-foot long stick*. Later that day, John was looking for something to wipe up a water spill on the kitchen floor. He saw the object that Jane had made and thought that it would be good for wiping up a water spill on the kitchen floor. He grabbed the object with the bundle of thick cloth pointing downward and pressed it against the water spill.
One day Jane wanted to wipe up a water spill on the kitchen floor [...]. *The object consisted of a bundle of plastic bags attached to a 4-foot long stick*. [...] pressed it against the water spill.

Figure 2: Examples of scenarios designed to evaluate the HIPE theory. Shown are the baseline and excerpted compromised structure scenarios with added emphasis.

Reimplementing Chaigneau et al.'s (2004) experiment on LLMs is motivated by several factors. First, given that the experiment they use involves text-only stimuli and responses, it can be comparably reimplemented on LLMs with little modification. Second, the CM hypothesized by HIPE theory is intuitive, straightforward, highly general, and relevant for many practical judgements about the physical world. This is unlike, for example, the more subtle theories concerning representation of natural kinds (Foster-Hanson & Rhodes, 2021). Furthermore, the qualitatively different results we obtain from GPT-4 on the one hand and GPT-3 and Falcon-40B on the other are made interesting by the fact that the common household terms studied here (such as "mop", "pencil", and "whistle") seem competently used even by GPT-3 (as we verify with a simple comprehension test). Thus, our results contribute to teasing apart representational capabilities that are quite similar at face value.

Experimental design

We replicate the crucial first two experiments from Chaigneau et al. (2004) on GPT-4 (gpt-4-031 version), GPT-3 (text-davinci-002 version), and Falcon-40B (all with temperature 0.7 and 5 max tokens). We investigate the extent to which compromising one of four aspects (goal, action, design history, or physical structure) of a scenario description impacts one of three outcomes (causality, function, or naming) across three artifact types (mop, pencil, or whistle).

First we focus on GPT-3 and consider multiple methods for serving the stimulus to it. The results that are reported were obtained using a setup that was most faithful to the one humans received, including warm-up trials and the possibility that answers to later questions could be influenced by subjects' exposure to earlier questions. Specifically, the scenarios are served to GPT-3 in a prompt that includes the same guidance that was given to the human participants by Chaigneau et al. (2004). The first element of the prompt is an introduction consisting of a description of the experiment

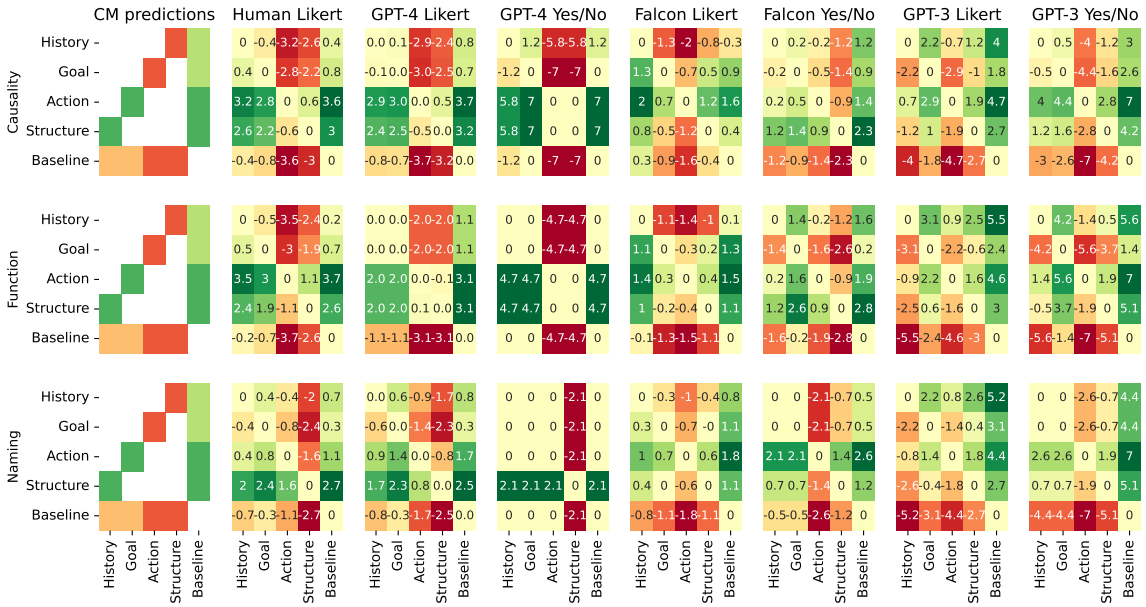

Figure 3: Heatmaps showing the pairwise difference between the scores when the factors on the x and y axes are compromised. Column one shows the predictions based on the CM hypothesized by HIPE: compromised scenarios should decrease outcome judgements relative to baseline, compromising action should be more significant than compromising goal, and compromising physical structure should be more significant than compromising design history. Column two shows the human data from Chaigneau et al. (2004), while columns three to eight show the data we obtain from GPT-4, Falcon-40B, and GPT-3. Columns three, five, and seven elicit responses from models using the same Likert scale as in the human experiment, whereas columns four, six, and eight use a Yes/No response.

and several demonstration responses. After the introduction the uncompromised scenario is presented and the Yes/No response is recorded. The compromised scenarios are then presented in random order. For each new scenario, we allow GPT-3 access to its entire response history as part of the prompt (since humans would have memory of their own past responses). The prompts we construct can thus be seen as a type of "in context learning". That is, when GPT-3 is generating a Yes/No reply for the $k+1$ scenario, it receives the introductory instructions and examples as well as the sequence of the first k scenarios and its own generated replies to those scenarios as part of the prompt. We record the probability that the model assigns to "yes" or versions thereof in response to each scenario. To control for possible effects from the order in which the compromised prompts are presented, two sets of results are averaged. Figure 2 shows excerpts from the prompts provided to GPT-3.

We also consider a further five variations on the above prompting design, which differ in how the scenarios are presented to GPT-3 and the manner in which a response is recorded. However, these alternative variations yield a lower Spearman correlation between the human and GPT-3 results (ranging between 0.28 and 0.5, versus 0.81 for the version presented here) and are less true to the original experiment. The largely negative result we will report for GPT-3 is strengthened by these numerous attempts to adapt the stimulus to it. By contrast, we will report positive results for GPT-4 despite only implementing the most faithful re-construction of the original experiment on it.

We reimplement the final version of the above experiment on GPT-4 following its release, and do the same with Falcon-40B. However, since OpenAI does not support the retrieval of probabilities associated with tokens generated by GPT-4, we instead repeat each question ten times and calculate the probability that the model generates a response including "yes" or versions thereof. The ten responses are split across two runs of five to control for the particular random order in which questions are presented. We report results in Figure 3 using the system message "You are a helpful assistant". Results from a version with the system message "You are a helpful assistant with an excellent understanding of the physical world" were also obtained (this slightly increases the correlation between the results from GPT-4 and human subjects, but is not necessary for observing a positive result). For parity we redo this experiment on GPT-3 with repeated output generation instead of the direct retrieval of output probabilities and present these results here. We find a Spearman correlation of 0.96 between the experiments run on GPT-3 with and without direct probability retrieval, giving us confidence that the results obtained from GPT-4 without direct access to probabilities are

comparable to what would be obtained with direct access. Because we find the results from GPT-4 to be more binary when using the Yes/No response as compared to the human data which was collected using a Likert scale, we re-implement the experiment with GPT-4 using a Likert response. For parity we do the same with GPT-3 and Falcon-40B.

We further investigate whether the models respond in the expected way to compromising distal factors by using a Likert scale re-implementation of Experiment 2 from Chaigneau et al. (2004), which tests a cumulative effect of compromising both distal factors while leaving the proximal factors unchanged. After providing the model with the same introductory prompt as above, we provide it with the baseline scenario, elicit its response on a Likert scale, accumulate this response, and then provide it successively with the next three scenarios in which one or both of function and history are compromised. We repeat this experiment for all three objects, running a given question / object combination twice with five responses collected from the model each time. We omit the naming question from this experiment for comparability with Chaigneau et al. (2004), who do the same. Full prompts and code are provided on Github.[1]

Results

Experiment 1

Figure 3 shows the predictions made according to the CM hypothesized by HIPE, the results obtained by Chaigneau et al. (2004) on human subjects, and the results obtained by us on GPT-4, GPT-3, and Falcon-40B. The CM predicts that history should have a less significant effect on outcome judgements than structure, and that goal should have a less significant effect than action. This corresponds to the diagonal of two green and two orange boxes in the CM predictions column of the figure. Furthermore, the CM predicts that compromising any factor should have a negative effect on outcome judgements relative to baseline, but with distal factors yielding a smaller negative effect than proximal factors. This corresponds to the orange horizontal and green vertical bars in the CM predictions column, lightening towards the left and top respectively due to the weaker effect of the distal factors.

The human subject results abide neatly by these predictions in the case of causality and function judgements. We observe a clear red box of four cells towards the top right, mirrored by a green box towards the bottom left. This subsumes the green/orange diagonal of the CM prediction, and corresponds to the stronger result of a larger effect of each proximal factor than both distal factors, rather than only a weaker result of structure being more significant than history and action being more significant than goal. We also see a green vertical bar on the right and a corresponding red bar on the bottom, lightening at the top and left respectively. This corresponds to every factor making a negative difference relative to baseline, with distal factors mattering less than proximal factors.

In the naming case, we see a somewhat different pattern in the human data than predicted by the CM. As the CM predicts, we observe a green column to the right that lightens in the top half, indicating that compromising any factor compromises the outcome judgement, but that proximal factors compromise the outcome judgement to a greater extent. However, a prominent red column in the fourth position mirrored by a green horizontal fourth row corresponds to a larger negative effect on the outcome when compromising structure than when compromising other features. This is intuitively reasonable: for example, using a bowl as a spoon does less to make it no longer be a bowl than flattening it does.

The results from GPT-4 bear a striking resemblance to the human data. In the causality and function heatmaps, we see a green vertical on the right and a red horizontal on the bottom, lightening towards the top right and bottom left respectively. We also see strong red boxes in the top right mirrored by green boxes in the bottom left. In the naming case, we see the same strong red column and green row appear in the fourth positions, corresponding to a dominating effect of compromising structure on the outcome compared to the effect of compromising other factors.

Overall there is a 0.88 Spearman correlation between the GPT-4 and human data with a Yes/No response (column four of Figure 3). The results from GPT-4 collected with a Yes/No response are more binary than the results from human subjects that were collected using a Likert scale response. In particular, the human data shows some effect of compromising distal factors thus not demonstrating full screening off of the distal factors by the proximal ones, while the Yes/No-response data from GPT-4 does not show this property. This is due to the difference in response modality, and we verify that re-implementing the experiment on GPT-4 with a Likert-scale response (column three of Figure 3) eliminates this effect and increases the Spearman correlation with the human data from 0.88 to 0.92.

The results from GPT-3 are less consistent with the predictions based on the CM and with the human data. First we focus on the results elicited using the same Likert scale from the human experiment (column seven of Figure 3). Across all three questions, we see that GPT-3 (like humans) consistently considers the compromised scenarios as less consistent with the concept than the baseline scenario. However, when comparing the effect of history to structure and the effect of goal to action across the three questions, GPT-3's responses only agree with the CM predictions in 3 out of 6 cases. Moreover, the Spearman correlation with the human data is only 0.67. However, these results may be due simply to the failure of GPT-3 to competently use a Likert scale. Indeed, after initially experimenting with a Likert scale on GPT-3, this approach was abandoned due to evidence that the model was not competently using such a response format. The results from a Likert scale experiment on GPT-3 are included here primarily for comparability with GPT-4, as the latter model appears to competently use the scale and exhibits the closest

[1] https://github.com/smusker/Causal_Models_Of_Word_Meaning

similarity to the human data when using this setup from the original human experiment.

When replacing the Likert scale with a simpler to use Yes/No response (column eight of Figure 3), the results from GPT-3 correlate more closely with the human data (Spearman correlation = 0.81). Additionally, when comparing the effect of history to structure and the effect of goal to action across the three questions, GPT-3's responses agree with the CM predictions in 5 out of 6 cases. However, these apparently positive observations cannot be taken at face value. First, as noted earlier, there is a high Spearman correlation of 0.96 between the version of the Yes/No experiment conducted on GPT-3 that uses the direct retrieval of the log probability of "yes" with the version that uses the frequency of "yes" generations shown here. However, in the former version, when comparing the effect of history to structure and the effect of goal to action across the three questions, GPT-3's responses only agree with the CM predictions in 4 out of 6 cases - closer to the chance level of 3/6. Second, as noted in the experimental design section, several reimplementations of the experiment on GPT-3 failed to yield positive results. Third, the human data show a pattern of high Spearman correlation between causality and function questions (0.99) with a much lower correlation between those questions and the naming one (0.64 causal/naming, 0.60 function/naming). Similarly, the GPT-4 data show Spearman correlations of 0.95 causal/function, 0.58 causal/naming, and 0.54 function/naming in the Yes/No response version. By contrast, in the GPT-3 Yes/No response data the correlation between questions is high in all comparisons (all pairwise correlations $\geq$ 0.89 in the Yes/No version and $\geq$ 0.85 in the Likert version). Furthermore, while in the naming question we see a very strong effect of structure compared to all other factors in the human and GPT-4 data, we see a stronger effect from action in the GPT-3 data (this can be seen in the redder fourth columns of the naming plots from the human and GPT-4 data, compared to the redder third column in the equivalent plot from the GPT-3 data). These discrepancies suggest nontrivial differences between how humans and GPT-4 on the one hand and GPT-3 on the other process these questions.

The results from Falcon-40B appear positive, although we will see in the following section that the model's performance in the second experiment casts doubt on the positivity of its results in the first experiment. Across the Likert (column five) and Yes/No (column 6) versions, when comparing the effect of history to structure and the effect of goal to action across the three questions, Falcon-40B's responses agree with the CM predictions in 12 out of 12 cases. With one exception (the effect of history in the case of causality), compromising any factor also has the expected negative effect relative to baseline. The correlation with human responses is moderate and is comparable to the correlation between GPT-3 and the human responses, at 0.74 in the Likert version and 0.62 in the Yes/No version. In the Yes/No version, the correlation between responses in different questions exhibits the same pattern as in the human and GPT-4 responses: 0.95 causal / function, 0.58 causal / naming, and 0.70 function / naming. However, the same pattern is not observed in the Likert version with Spearman correlations of 0.94 causal / function, 0.79 causal / naming, and 0.90 function / naming (the latter correlation in particular is expected to be low but is not).

Experiment 2

Experiment 1 primarily tests that the subject exhibits the screening off of distal factors: i.e., when distal factors are compromised but the proximal factors that are hypothesized to mediate their effect are left unchanged, the compromising effect of the distal factors should be largely masked.

Nevertheless, one should expect compromising the distal factors to have some effect on outcome judgements and indeed this is observed in Experiment 1. Experiment 2 further investigates the effect of compromising distal factors by verifying that compromising each distal factor independently results in a lowered outcome judgement relative to baseline and that compromising both of these factors together results in an even lower outcome judgement. Following Chaigneau et al. (2004) for comparability, we present results from LLMs that average function and causal outcome judgements across the three object types.

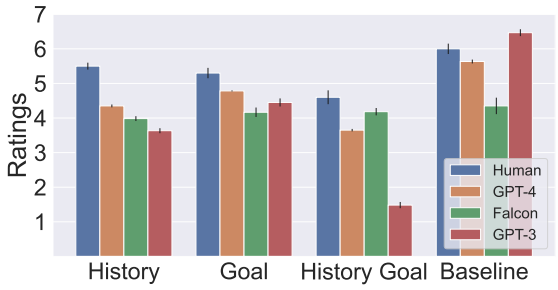

Figure 4: Human and LLM data collected in Experiment 2. Human data are from Chaigneau et al. (2004). The chart shows the subject's mean rating, averaged across function and causality for all three objects, when the factor or factors shown on the x-axis are compromised. Error bars are the Standard Error of the Mean.

As can be observed in Figure 4, data collected from GPT-4 and -3 follow the same trend observed in the human subjects[2]. In particular, compromising history, goal, or both together each yield a lower outcome rating than baseline. Moreover, compromising both of these distal factors together yields a

[2]Note that human subjects were prompted with "warm-up" questions using the Likert scale, which they are encouraged to discuss with each other. We have access to these questions but not records of human responses, and so we provide GPT-4 with these questions accompanied with our own responses to them as part of its prompt. This successfully primes GPT-4 and -3 to use the scale meaningfully, but it should not be expected to be calibrated to the absolute magnitude of the original human subjects.

Comparisons (GPT-4)	t value	Significant?
History < Baseline	18.6	Yes
Goal < Baseline	14.7	Yes
History Goal < Baseline	30.1	Yes
History > History Goal	13.0	Yes
Goal > History Goal	29.8	Yes
Comparisons (GPT-3)	**t value**	**Significant?**
History < Baseline	23.0	Yes
Goal < Baseline	13.1	Yes
History Goal < Baseline	37.1	Yes
History > History Goal	18.6	Yes
Goal > History Goal	20.1	Yes
Comparisons (Falcon)	**t value**	**Significant?**
History < Baseline	1.5	No
Goal < Baseline	0.7	No
History Goal < Baseline	0.6	No
History > History Goal	NA	NA
Goal > History Goal	0.1	No

Table 1: Table showing details from statistical testing of the difference between mean response values across the questions of Experiment 2. Chaigneau et al. (2004) conduct the same comparisons using the human data, finding that the expected differences are significant to at least the $p \leq 0.01$ level. We test for significance at the 0.05 level, and report NA where the difference in value is not in the predicted direction thus making statistical testing unnecessary. In all cases of significance at the 0.05 level, we also observe significance to at the least the $p \leq 0.0005$ level. Our data for the statistical testing are the responses from LLMs shown in Figure 4.

lower outcome rating than compromising either of them independently. Chaigneau et al. (2004) find that these five pairwise comparisons are statistically significant in the human data, and we find that the same holds in the data collected from GPT-4 and -3. By contrast, Falcon-40B fails to produce the expected results. The combined effect of history and goal is not observed to be greater than the separate effects of those factors, yielding a negative result for Falcon-40B on Experiment 2. Moreover, in Experiment 2 we do not find history and goal independently to have a negative effect relative to baseline. This constitutes a failure of Falcon-40B on a subset of the comparisons of interest from Experiment 1, which weakens the fairly positive results from that model in the earlier experiment. Further information is shown in Table 1.

Discussion and limitations

Our results show a similarity between the responses of GPT-4 and human subjects in both Experiment 1 and 2. We observe positive results with GPT-3 on Experiment 2, but mixed results on Experiment 1. Falcon-40B shows relevant similarity to the human responses in Experiment 1 but not in Experiment 2, and it fails in Experiment 2 in a manner that is not consistent with success on Experiment 1. Overall our results show a marked difference between the responses from human subjects and GPT-4 on the one hand and GPT-3 / Falcon-40B on the other, which may suggest a qualitative difference between these models in how they represent common artifacts.

At the highest level, we interpret these results as speaking to the need for a broad and rigorous discussion about evaluation in the modern age of LLMs. Recent years have relied increasingly on tests from cognitive science and psycholinguistics as a source of more rigorous, more controlled, and more hypothesis-driven evaluations of language models (Bastings et al., 2022). Such experiments have been primarily fruitful in the context of two types of arguments. First, they have produced insightful negative results (e.g., Ettinger (2020)). In such cases, models' failure on psycholinguistic tests can be taken as evidence that the models probably lack at least *some* aspect of whatever mechanism humans use to perform the same tasks. Second, such tests have produced insightful positive results (e.g., Linzen et al. (2016)). For example, models' success has been used specifically to counter learnability or "poverty of the stimulus" arguments, and thus to question the usefulness of specific diagnostic tests. That is, if some behavior is assumed to require a given capacity, and a model that is known to lack that capacity nonetheless produces that behavior, then a different test is needed to diagnose the capacity of interest.

The present study may best be viewed as an instance of the latter. Theories like HIPE are generally assumed to be tests of causal models which presuppose that agents' representations are grounded in the physical and goal-oriented world. If models with access only to text (or at most text and images) are presumed to lack this grounding, then the success of models on this task may suggest that the human results on HIPE tests are not *necessarily* diagnostic of such grounding. Thus, further tests must be developed to determine what representations underlie models' (and humans') behavior.

Importantly, caution should be exercised in interpreting positive results on psychological tests as diagnostic of "human-like" or even "human-level" processing. Only in some cases are positive results in such tests clearly interpretable, such as against a backdrop of a clear learnability argument (i.e., a claim about some capacity that the model being studied is known *a priori* not to possess). However, the likelihood of increased positive results in the LLM era could lead to psycholinguistic tests being hastily viewed as diagnostic of human-like processing. Thus, we raise questions about what role such tests should play in future evaluations of models' representations. In particular, if success on behavioral tests alone is not a sufficient test of competence, what is?

Our work is limited in that we use the same materials as Chaigneau et al. (2004) and aim to preserve comparability with the data they collect from human subjects. As such, we consider only the three artifacts from the original study and average results across them. Results could differ if more objects were included. However, comparable data from human subjects do not appear to exist for a broader class of objects.

Bibliographical References

References

Bastings, J., Belinkov, Y., Elazar, Y., Hupkes, D., Saphra, N., & Wiegreffe, S. (Eds.). (2022, December). *Proceedings of the fifth blackboxnlp workshop on analyzing and interpreting neural networks for nlp.* Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.blackboxnlp-1.0

Bisk, Y., Zellers, R., Le bras, R., Gao, J., & Choi, Y. (2020, Apr.). Piqa: Reasoning about physical commonsense in natural language. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(05), 7432-7439. Retrieved from https://ojs.aaai.org/index.php/AAAI/article/view/6239 doi: 10.1609/aaai.v34i05.6239

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... others (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Carey, S. (2009). *The origin of concepts*. Oxford University Press. Retrieved from https://books.google.com/books?id=h9J2DwAAQBAJ

Chaigneau, S. E., Barsalou, L. W., & Sloman, S. A. (2004). Assessing the causal structure of function. *Journal of Experimental Psychology: General*, *133*(4), 601–225. doi: 10.1037/0096-3445.133.4.601

Da, J., & Kasai, J. (2019). Cracking the contextual commonsense code: Understanding commonsense reasoning aptitude of deep contextual representations. In *Proceedings of the first workshop on commonsense inference in natural language processing* (pp. 1–12).

Ettinger, A. (2020). What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, *8*, 34–48.

Forbes, M., Holtzman, A., & Choi, Y. (2019). *Do neural language representations learn physical commonsense?* arXiv. Retrieved from https://arxiv.org/abs/1908.02899 doi: 10.48550/ARXIV.1908.02899

Foster-Hanson, E., & Rhodes, M. (2021). The psychology of natural kind terms. In S. Biggs & H. Geirsson (Eds.), *The routledge handbook of linguistic reference* (p. 295-308). New York: Routledge.

Futrell, R., Wilcox, E., Morita, T., Qian, P., Ballesteros, M., & Levy, R. (2019). Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 32–42).

Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., & Lee, S. (2019, 09–15 Jun). Counterfactual visual explanations. In K. Chaudhuri & R. Salakhutdinov (Eds.), *Proceedings of the 36th international conference on machine learning* (Vol. 97, pp. 2376–2384). PMLR. Retrieved from https://proceedings.mlr.press/v97/goyal19a.html

Greenberg, M., & Harman, G. (2005). Conceptual role semantics. In E. Lepore & B. C. Smith (Eds.), *The oxford handbook of philosophy of language* (p. 295-322). Oxford: Oxford University Press.

Huang, P.-S., Zhang, H., Jiang, R., Stanforth, R., Welbl, J., Rae, J., ... Kohli, P. (2020, November). Reducing sentiment bias in language models via counterfactual evaluation. In *Findings of the association for computational linguistics: Emnlp 2020* (pp. 65–83). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.findings-emnlp.7 doi: 10.18653/v1/2020.findings-emnlp.7

Huang, W., Abbeel, P., Pathak, D., & Mordatch, I. (2022). *Language models as zero-shot planners: Extracting actionable knowledge for embodied agents.* arXiv. Retrieved from https://arxiv.org/abs/2201.07207 doi: 10.48550/ARXIV.2201.07207

Kassner, N., & Schütze, H. (2020). Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 7811–7818).

Keil, F. C. (1989). *Concepts, kinds and development*. MIT Press.

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, *4*, 521–535. Retrieved from https://aclanthology.org/Q16-1037 doi: 10.1162/tacl_a_00115

Marvin, R., & Linzen, T. (2018). Targeted syntactic evaluation of language models. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 1192–1202).

OpenAI. (2023). *Gpt-4 technical report.*

Park, J., & Sloman, S. A. (2016). Causal models and screening-off. In *A companion to experimental philosophy* (p. 450-462). John Wiley & Sons, Ltd. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118661666.ch31 doi: https://doi.org/10.1002/9781118661666.ch31

Pavlick, E. (2022). Semantic structure in deep learning. *Annual Review of Linguistics*, *8*, 447–471.

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., & Miller, A. (2019). Language models as knowledge bases? In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)* (pp. 2463–2473).

Sloman, S. (2005). *Causal models: How people think about the world and its alternatives*. Oxford University Press. Retrieved from https://books.google.com/books?id=d9GTO0rl58QC

Technology Innovation Institute. (2023). *Falcon language model.* https://falconllm.tii.ae/. (Accessed: 15 October 2023)

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., & Bowman, S. R. (2020). Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, *8*, 377–392.